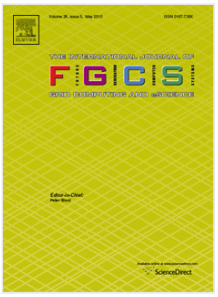
Accepted Manuscript



A family of enhanced (L, α)-diversity models for privacy preserving data publishing

Xiaoxun Sun, Min Li, Hua Wang

PII: S0167-739X(10)00141-X
DOI: 10.1016/j.future.2010.07.007
Reference: FUTURE 1916

To appear in: *Future Generation Computer Systems*

Received date: 22 December 2009
Revised date: 9 May 2010
Accepted date: 18 July 2010

Please cite this article as: X. Sun, M. Li, H. Wang, A family of enhanced (L, α)-diversity models for privacy preserving data publishing, *Future Generation Computer Systems* (2010), doi:10.1016/j.future.2010.07.007

A Family of Enhanced (L, α)-Diversity Models For Privacy Preserving Data Publishing

Xiaoxun Sun, Min Li, Hua Wang

Department of Mathematics & Computing

University of Southern Queensland, Australia

Email: {sunx, limin, wang}@usq.edu.au

Abstract

Privacy preservation is an important issue in the release of data for mining purposes. Recently, a novel l-diversity privacy model was proposed, however, even an l-diverse data set may have some severe problems leading to reveal individual sensitive information. In this paper, we remedy the problem by introducing distinct (l, α)-diversity, which, intuitively, demands that the total weight of the sensitive values in a given QI-group is at least α, where the weight is controlled by a pre-defined recursive metric system. We provide a thorough analysis of the distinct (l, α)-diversity and prove that the optimal distinct (l, α)-diversity problem with its two variants entropy (l, α)-diversity and recursive (c, l, α)-diversity are NP-hard, and propose a top-down anonymization approach to solve the distinct (l, α)-diversity problem with its variants. We show in the extensive experimental evaluations that the proposed methods are practical in terms of utility measurements and can be implemented efficiently.

1 Introduction

Many data holders publish their microdata for different purposes. However, they have difficulties in releasing information such that no privacy is compromised. The traditional approach of releasing the data tables without breaching the privacy of individuals in the table is to de-identify records by removing the identifying fields such as name, address, and social security number. However, joining this de-identified table with a publicly available database (like the voters database) on attributes like race, age, and zip code (usually called quasi-identifier)

can be used to identify individuals. For example, Sweeney reported in [31] that 87% of the population of the United States can be uniquely identified by the combinations of attributes: gender, date of birth, and 5-digit zip code.

In order to protect privacy, Sweeney [31] proposed the k-anonymity model, where some of the quasi-identifier fields are suppressed or generalized so that, for each record in the modified table, there are at least $(k-1)$ other records in the modified table that are identical to it along the quasi-identifier attributes. In the literature of k-anonymity problem, there are two main models. One model is global recoding [9, 14, 26, 30] while the other is local recoding [2, 30]. Here, we assume that each attribute has a corresponding conceptual generalization hierarchy or taxonomy tree. A lower level domain in the hierarchy provides more details than a higher level domain. For example, Zip Code 14248 is a lower level domain and Zip Code 142** is a higher level domain. We assume such hierarchies for numerical attributes too. In particular, we have a hierarchical structure defined with {value, interval, $*$}, where value is the raw numerical data, interval is the range of the raw data and $*$ is a symbol representing any values. Generalization replaces lower level domain values with higher level domain values. For example, Age 27, 28 in the lower level can be replaced by the interval (27-28) in the higher level.

1.1 Motivation

When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in the literature [4, 13]: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure happens when the new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before releasing the data. Although k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. Several models such as p-sensitive k-anonymity [32], l-diversity [23] and t-closeness [18] were proposed. However, depending on the nature of the sensitive attributes, even these enhanced properties still permits the information to be disclosed.

p-sensitive k-anonymity principle: The purpose of p-sensitive k-anonymity is to protect against attribute disclosure by requiring that there be at least p different values for each sensitive attribute within the records sharing a combination of quasi-identifier. This approach has the limitation of implicitly assuming that each sensitive attribute takes values uniformly over its domain; that is, that the frequencies of the various values of a sensitive attribute are similar. When this is not the case, achieving the required level of privacy may cause a huge data

utility loss.

l-diversity principle: The l-diversity model protects against sensitive attribute disclosure by considering the distribution of the attributes. The approach requires l "well-represented"[1] values in each combination of quasi-identifiers. This may be difficult to achieve and, like p-sensitive k-anonymity, may result in a large data utility loss. Further, as we shall discuss in Section 2, l-diversity is insufficient to prevent similarity attack.

t-closeness principle: The t-closeness model protects against sensitive attributes disclosure by defining semantic distance among sensitive attributes. The approach requires the distance between the distribution of the sensitive attribute in the group and the distribution of the attribute in the whole data set to be no more than a threshold t. Whereas Li et al. [18] elaborate on several ways to check t-closeness, no computational procedure to enforce this property is given. If such a procedure was available, it would greatly damage the utility of data because enforcing t-closeness destroys the correlations between quasi-identifier attributes and sensitive attributes.

Faced with these limitations, we intend to enhance the current privacy paradigms to make them preserve the better trade-off between data quality and privacy. The work presented in this paper is highly inspired by [23]. The main contribution of [23] is to introduce the basic l-diversity property, which provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary. In this paper, we propose a family of enhanced (l, α)-diversity models, where l is an integer and α is a real number. In addition to l-diversity, we further require that the total weight of sensitive values in any QI-group should be at least α after modification. We also propose an efficient anonymization method to tackle our problems.

2 Preliminaries

Let T be the initial microdata and T' be the released microdata. T' consists of a set of tuples over an attribute set. The attributes characterizing microdata are classified into the following three categories.

- *Identifier attributes* can be used to identify a record such as Name and Medicare card.
- *Quasi-identifier (QI) attributes* may be known by an intruder, such as Zip code and Age. QI attributes are presented in the released microdata T' as well as in T.

[1]The interpretation of the term "well-represented" can be found in [23].

ID	Age	Country	Zip Code	Disease
1	27	USA	14248	HIV
2	28	Canada	14207	HIV
3	26	USA	14206	Cancer
4	25	Canada	14249	Cancer
5	41	China	13053	Hepatitis
6	48	Japan	13074	Phthisis
7	45	India	13064	Asthma
8	42	India	13062	Obesity
9	33	USA	14242	Flu
10	37	Canada	14204	Flu
11	36	Canada	14205	Flu
12	35	USA	14248	Indigestion

Table 1: The raw microdata

ID	Age	Country	Zip Code	Disease
1	(27-28)	America	142**	HIV
2	(27-28)	America	142**	HIV
3	(25-26)	America	142**	Cancer
4	(25-26)	America	142**	Cancer
5	>40	Asia	130**	Hepatitis
6	>40	Asia	130**	Phthisis
7	>40	Asia	130**	Asthma
8	>40	Asia	130**	Obesity
9	(33-35)	America	142**	Flu
10	(36-37)	America	142**	Flu
11	(36-37)	America	142**	Flu
12	(33-35)	America	142**	Indigestion

Table 2: 2-anonymous microdata

- *Sensitive attributes* are assumed to be unknown to an intruder and need to be protected, such as Disease or ICD-9 Code[2]. Sensitive attributes are presented both in T and T'.

In what follows we assume that the identifier attributes have been removed and the quasi-identifier and sensitive attributes are usually kept in the released and initial microdata table. Another assumption is that the values of the sensitive attributes are not available from any external source. This assumption guarantees that an intruder can not use the sensitive attributes to increase the chances of disclosure. Unfortunately, an intruder may use record linkage techniques [35] between quasi-identifier attributes and external available information to glean the identity of individuals from the modified microdata. To avoid this possibility of privacy disclosure, one frequently used solution is to modify the initial microdata, more specifically the quasi-identifier attributes values (a minimal set Q of attributes in T that can be joined with external information to re-identify individual records), in order to enforce the k-anonymity property.

Definition 1 (k-anonymity) *T' is said to satisfy k-anonymity if and only if each combination of quasi-identifier attributes in T' occurs at least k times.*

A QI-group in the modified microdata T' is the set of all records in the table containing identical values for the QI attributes. There is no consensus in the literature over the term

[2]International Statistical Classification of Diseases and Related Health Problems: ICD-9 -provides multiple external links for looking up ICD codes. Available `http://icd9cm.chrisendres.com/`.

Name	Age	Country	Zip Code
Rick	26	USA	14246
Hassen	45	India	13064
Rudy	25	Canada	14249
Yamazaki	48	Japan	13074

Table 3: External available information

Category ID	Sensitive Information	Sensitivity
One	HIV, Cancer	Top Secret
Two	Phthisis, Hepatitis	Secret
Three	Obesity, Asthma	Less Secret
Four	Flu, Indigestion	Non Secret

Table 4: Categories of Disease

ID	Age	Country	Zip Code	Disease
1	<30	America	142**	HIV
2	<30	America	142**	HIV
3	<30	America	142**	Cancer
4	<30	America	142**	Cancer
5	>40	Asia	130**	Hepatitis
6	>40	Asia	130**	Phthisis
7	>40	Asia	130**	Asthma
8	>40	Asia	130**	Obesity
9	3*	America	142**	Flu
10	3*	America	142**	Flu
11	3*	America	142**	Flu
12	3*	America	142**	Indigestion

Table 5: 2-diverse microdata

used to denote a QI-group. This term was not defined when k-anonymity was introduced [26, 31]. More recent papers use different terminologies such as equivalence class [36] and QI-cluster [33].

For example, let the set {Age, Country, Zip Code} be the quasi-identifier of Table 1. Table 2 is one 2-anonymous view of Table 1 since there are five QI-groups and the size of each QI-group is at least 2. So k-anonymity can ensure that even though an intruder knows a particular individual is in the k-anonymous microdata table T, s/he can not infer which record in T corresponds to the individual with a probability greater than $1/k$.

The k-anonymity property ensures protection against identity disclosure, i.e. the identification of an entity (person, institution). However, as we will show next, it does not protect the data against attribute disclosure, which occurs when the intruder finds something new about a target entity. Consider Table 2, where the set of quasi-identifiers is composed of {Age, Country, Zip Code} and Disease is the sensitive attribute. As we discussed above, identity disclosure does not happen in this modified microdata. However, assuming that external information in Table 3 is available, attribute disclosure can take place. If the intruder knows that in Table 2 the Age attribute was modified to '(25-26)', s/he can deduce that both Rick and Rudy have Cancer, even he does not know which record, 3 or 4, is corresponding to which person. This example shows that even if k-anonymity can protect identity disclosure, sometimes it fails to protect against sensitive attribute disclosure. To deal with this problem in privacy breach, the l-diveristy model was introduced in [23].

Definition 2 (l-diversity) *A QI-group is said to have l-diversity if there are at least l dis-*

tinct values for the sensitive attribute. A modified table is said to have l-diversity if every QI-group of the table has l-diversity.

For instance, Table 5 is a 2-diverse view of Table 1. Although the l-diversity principle represents an important step beyond k-anonymity in protecting sensitive attribute disclosures, it still has some shortcomings. Following through, we show that the l-diversity principle is insufficient to prevent the *similarity attack*, which means when the sensitive attribute values in a QI-group are distinct but with similar sensitivity, an adversary can learn important information.

Sometimes, the domain of the sensitive attributes, especially the categorical ones, can be partitioned into categories according to the sensitivity of attributes. For example, in the medical data set Table 1, the Disease attribute can be classified into four categories (see Table 4). The different types of diseases are organized in a category domain. The attribute values are very specific, for example they can represent HIV or Cancer, which are both Top Secret information of the individuals. In the case that the initial microdata contains specific sensitive attributes like Disease, the data owner can be interested in protecting not only these most specific values, but also the category that the sensitive values belong to. For example, the information of a person who affected with Top Secret needs to be protected, no matter whether it is HIV or Cancer. If we modify the microdata to just satisfy l-diversity property, it is possible that in a QI-group with l distinct sensitive attribute values, all of them belong to the same pre-defined confidential category. For instance, the values {HIV, HIV, Cancer, Cancer} of one QI-group in Table 5 all belong to Top Secret category. To avoid such situations, we introduce a family of enhanced (l, α)-diversity model integrating a recursive metric function.

3 A family of enhanced (l, α)-diversity models

Let S be a categorical sensitive attribute we want to protect against attribute disclosure. First, we sort the values of S according to their sensitivity, forming an ordered value domain D, and then partition the attribute domain into m-categories $(S_1, S_2, \cdots, S_m)$, such that $S = \cup_{i=1}^{m} S_i$, $S_i \cap S_j = \emptyset$ (for $i \neq j$) and $S_l \leq S_k$ $(1 \leq l \leq k)$. We say $S_l \leq S_k$, if S_l is more sensitive than S_k $(1 \leq l \leq k)$. For example, consider the Disease S={HIV, Cancer, Phthisis, Hepatitis, Obesity, Asthma, Flu, Indigestion} in Table 1, it has been partitioned into four categories according to the sensitivity of the diseases (Table 4), where S_1 (Top Secret) is the most sensitive and S_4 (Non Secret) is the least one.

Definition 3 *Let $D(S) = \{S_1, S_2, \cdots, S_k\}$ denote a partition of categorical domain of an attribute S, weight(S_i) be the weight of category S_i and $w_{i,i-1}$ be the weight between two*

Age	Country	Zip Code	Disease
<40	America	142**	HIV
<40	America	142**	HIV
<40	America	142**	Cancer
<40	America	142**	Flu
>40	Asia	130**	Hepatitis
>40	Asia	130**	Phthisis
>40	Asia	130**	Asthma
>40	Asia	130**	Obesity
<40	America	14***	Cancer
<40	America	14***	Flu
<40	America	14***	Flu
<40	America	14***	Indigestion

Table 6: Distinct (3,1)-diversity

Age	Country	Zip Code	Disease
<40	America	142**	HIV
<40	America	142**	HIV
<40	America	142**	Flu
<40	America	142**	Flu
>40	Asia	130**	Hepatitis
>40	Asia	130**	Phthisis
>40	Asia	130**	Asthma
>40	Asia	130**	Obesity
<40	America	14***	Cancer
<40	America	14***	Cancer
<40	America	14***	Flu
<40	America	14***	Indigestion

Table 7: Entropy (2,2)-diversity

adjacent categories ($2 \leq i \leq k$). Then,

$$\frac{weight(S_{i+1}) - weight(S_i)}{weight(S_i) - weight(S_{i-1})} = \frac{w_{i+1,i}}{w_{i,i-1}}, 2 \leq i \leq k \tag{1}$$

Where $weight(S_1) = 0$, $weight(S_k) = 1$ is the initial condition. Note that the weight of the specific sensitive value is equal to the weight of the category that the specific value belongs to. The weight of the QI-group is the total weight of each specific sensitive value that the QI-group contains. In the following, we discuss two simple and typical schemes to define $w_{i,i-1}$.

(1): uniform weight: $w_{i,i-1} = 1$ ($2 \leq i \leq k$). This is the simplest scheme when all weights among the categories are equal to 1. In this scheme, the weight of the category S_i is the number of categories that are less sensitive than S_i over the total number of categories. For example, given the partition of sensitive attributes as shown in Table 4 and A={Cancer, Phthisis, Asthma, Flu}. The distance between Cancer (S_1) and Flu (S_4) is 3/3=1, while the distance between Phthisis (S_2) and Asthma (S_3) is 1/3. According to Equation (1), $weight(S_1) = 0$, $weight(S_2) = 1/3$ and $weight(Asthma) = 2/3$, $weight(Flu) = 1$, the total weight of A is 0+1/3+2/3+1=2.

(2): sensitivity weight: $w_{i,i-1} = \frac{1}{(i-1)^\beta}$ ($2 \leq i \leq k$, $\beta \geq 1$). For a fixed β, the intuition of this scheme is that the weight of more confidential categories should possess less weight than the less ones. Thus, we formulate the sensitivity weight scheme, where the weight near to the top confidential category is smaller and the weight far from the top is larger. Still consider the partition of sensitive attributes as shown in Table 4 and A={Cancer, Phthisis, Asthma,

Flu}. When setting $\beta = 2$ and according to Equation (1), $weight(S_1) = 0$, $weight(S_2) = 9/61$ and $weight(Asthma) = 45/61$, $weight(Flu) = 1$, the total weight of A is 115/61.

Definition 4 (distinct (l, α)-diversity) *The modified microdata T' satisfies distinct (l, α)-diversity if it satisfies l-diversity principle, and for each QI-group, the total weight of its sensitive attribute values is at least α.*

Table 6 is a distinct $(3, 1)$-diverse view of Table 1. Since there are at least three different values in each QI-group and the least total weight of the QI-group is 1. Compared with Table 5, we can easily see that requiring the distinct (l, α)-diversity can significantly reduce the risk of sensitive attribute disclosure, hence better protecting individual's private information. Further, if we take a closer look at the first four tuples in Table 6, which form a QI-group, and we assume that the attacker has some background knowledge that Allen falls in that group, then although the attacker does not have 100% confidence to say that Allen suffered from deadly disease, he/she still has a higher probability of 75% to infer the sensitive information of Allen. In order to avoid this situation, we further introduce two more variants of distinct (l, α)-diversity models, which taking the amount of sensitive information into account.

Definition 5 (entropy (l, α)-diversity) *Let the entropy of a QI-group G be defined as: $Entropy(G) = -\sum_{s \in S} p(G, s) logp(G, s)$, in which S is the set of the categories divided among the sensitive attribute values, and $p(G, s)$ is the fraction of records in G that have sensitive value s in the category S. The modified microdata table T' satisfies entropy (l, α)-diversity if it satisfies (l, α)-diversity principle, and for every QI-group G, $Entropy(G) \geq log(l)$.*

Table 7 is a distinct $(3, 1)$-diverse view of Table 1. The entropy (l, α)-diversity principle is stronger than the distinct (l, α)-diversity. In order to have entropy (l, α)-diversity for each QI-group, the entropy of the entire table must be at least $log(l)$. Sometimes this may be too restrictive, since the entropy of the entire table may be low if a few values are very common. This leads to the following less conservative variant of (l, α)-diversity models.

Let m be the number of categories of the sensitive attribute values belong to in a QI-group, and r_i, $1 \leq i \leq m$ be the number of times that the i^{th} most frequent category appears in a QI-group G. G is said to have recursive (c, l, α)-diversity if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$.

Definition 6 (recursive (c, l, α)-diversity) *The modified microdata T' satisfies recursive (c, l, α)-diversity if it satisfies (l, α)-diversity principle, and all of its QI-groups have recursive (c, l, α)-diversity.*

The recursive (c, l, α)-diversity ensures that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely, which, intuitively, could balance the distribution of non-sensitive and sensitive attributes in each QI-group.

Among these variants of (l, α)-diversity models, both parameters l and α are intuitive and operable in real-world applications. Parameter l specifies the number of "well-represented" values in each QI-group, while parameter α captures the degree each specific sensitive attribute value contributes to the QI-group. By increasing the value of α or l, we are strengthening the protection from sensitive attribute disclosure, however, in different ways. Specifically, the effect of raising α is to enlarge the protection range of each sensitive value, whereas the purpose of elevating l is to lower an adversarys chance of beating that protection. The enhanced (l, α)-diversity models have the following monotonicity property.

Proposition 1 *Given a data set T and α, if T satisfies distinct (l_1, α)-diversity, it also satisfies (l_2, α)-diversity, for every $l_2 \leq l_1$.*

Following through, we define the optimal problem of distinct (l, α)-diversity, and prove that the optimal distinct (l, α)-diversity is NP-hard, and as a corollary, we deduce that both the optimal entropy (l, α)-diversity and recursive (c, l, α)-diversity problems are also NP-hard.

Theorem 1 *: The optimal distinct (l, α)-diversity problem is NP-hard for a binary alphabet ($\sum = \{0, 1\}$).*

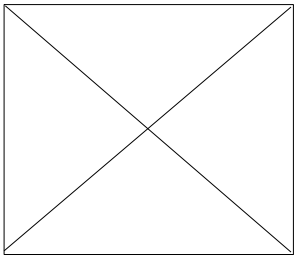
Proof: The proof is by transforming the problem of Edge Partition into 4-Cliques [10] to the distinct (l, α)-diversity problem.

Edge Partition into 4-Cliques : Given a simple graph $G = (V, E)$, with $|E| = 6m$ for some integer m, can the edges of G be partitioned into m edge-disjoint 4-cliques?

Given an instance of Edge Partition into 4-Cliques. Set $l = 2$, $\alpha = 6$. We construct the data set T as follows: for each edge $e = (v_1, v_2) \in E$, create a pair of records $r_{v1,v2}$ and $\tilde{r}_{v1,v2}$, so there are $2|E|$ in total. For each vertex $v \in V$, construct a non-sensitive attribute, there are totally $|V|$ non-sensitive attributes. In addition to it, we create one sensitive attribute in T, which makes the number of attributes in T be $|V| + 1$. The two records $(r_{v_i,v_j}, \tilde{r}_{v_i,v_j})$ have the attribute values of both v_i and v_j equal to 1 and all other non-sensitive attribute values equal to 0, but one record r_{v_i,v_j} has the sensitive attribute equal to 1 and the other record $\tilde{r}_{v_i,v_j}$ has the sensitive attribute equal to 0 $(1 \leq i, j \leq |V|)$. An example is given in Figure 1, where the data set T in Figure 1(b) is constructed from the clique of Figure 1(a).

We define the cost of the distinct (2,6)-diversity to be the number of suppressions applied in the data set. We show that the cost of the distinct (2,6)-diversity is at most $48m$ if and only if E can be partitioned into a collection of m edge-disjoint 4-cliques.

"$\Leftarrow$" Suppose E can be partitioned into a collection of m disjoint 4-cliques. Consider a 4-clique C with vertices v_1, v_2, v_3 and v_4. If we suppress the attributes v_1, v_2, v_3 and v_4 in the 12 records corresponding to the edges in C, then a cluster of these 12 records are formed where each modified record has four $*$'s. Note that the distinct (l, α)-diversity requirement



ID	A_1	A_2	A_3	A_4	S
r_{v_1,v_2}	1	1	0	0	1
$\bar{r}_{v_1,v_2}$	1	1	0	0	0
r_{v_1,v_3}	1	0	1	0	1
$\bar{r}_{v_1,v_3}$	1	0	1	0	0
r_{v_1,v_4}	1	0	0	1	1
$\bar{r}_{v_1,v_4}$	1	0	0	1	0
r_{v_2,v_3}	0	1	1	0	1
$\bar{r}_{v_2,v_3}$	0	1	1	0	0
r_{v_2,v_4}	0	1	0	1	1
$\bar{r}_{v_2,v_4}$	0	1	0	1	0
r_{v_3,v_4}	0	0	1	1	1
$\bar{r}_{v_3,v_4}$	0	0	1	1	0

Figure 1: (a) one 4-clique C; (b) a data set T constructed from C

can be satisfied as the frequency of the sensitive attribute value 1 is equal to 6 and the distinct number of sensitive values are 2. The cost of the distinct (2,6)-diversity is equal to $12 \times 4 \times m = 48m$. An example of the anonymization is given in Figure 2, where Figure 2(b) is a (2,6)-diverse view of Figure 2(a).

ID	A_1	A_2	A_3	A_4	S
r_{v_1,v_2}	1	1	0	0	1
$\bar{r}_{v_1,v_2}$	1	1	0	0	0
r_{v_1,v_3}	1	0	1	0	1
$\bar{r}_{v_1,v_3}$	1	0	1	0	0
r_{v_1,v_4}	1	0	0	1	1
$\bar{r}_{v_1,v_4}$	1	0	0	1	0
r_{v_2,v_3}	0	1	1	0	1
$\bar{r}_{v_2,v_3}$	0	1	1	0	0
r_{v_2,v_4}	0	1	0	1	1
$\bar{r}_{v_2,v_4}$	0	1	0	1	0
r_{v_3,v_4}	0	0	1	1	1
$\bar{r}_{v_3,v_4}$	0	0	1	1	0

ID	A_1	A_2	A_3	A_4	S
r_{v_1,v_2}	*	*	*	*	1
$\bar{r}_{v_1,v_2}$	*	*	*	*	0
r_{v_1,v_3}	*	*	*	*	1
$\bar{r}_{v_1,v_3}$	*	*	*	*	0
r_{v_1,v_4}	*	*	*	*	1
$\bar{r}_{v_1,v_4}$	*	*	*	*	0
r_{v_2,v_3}	*	*	*	*	1
$\bar{r}_{v_2,v_3}$	*	*	*	*	0
r_{v_2,v_4}	*	*	*	*	1
$\bar{r}_{v_2,v_4}$	*	*	*	*	0
r_{v_3,v_4}	*	*	*	*	1
$\bar{r}_{v_3,v_4}$	*	*	*	*	0

Figure 2: (a) an original data set; (b) a distinct (2,6)-diverse data set

"⇒" Suppose the cost of the distinct (2,6)-diversity is at most $48m$. As G is a simple graph, any twelve records should have at least four attributes different. So, each record should have at least four $*$s in the solution of the distinct (2,6)-diversity. Then, the cost of the distinct (2,6)-diversity is at least $12 \times 4 \times m = 48m$. Combining with the proposition that the cost is at most $48m$, we obtain the cost is exactly equal to $48m$ and thus each record should have exactly four $*$'s in the solution. Each cluster should have exactly 12 records (where six have sensitive value 1 and the other six have sensitive value 0). Suppose the twelve modified records contain four $*$'s in attributes v_1, v_2, v_3 and v_4, the records contain 0s in all other nonsensitive attributes. This corresponds to a 4-clique with vertices v_1, v_2, v_3 and v_4. Thus, we conclude that the solution corresponds to a partition into a collection of m edge-disjoint 4-cliques. ■

Corollary 1 *: Both the optimal entropy (l, α)-diversity and recursive (c, l, α)-diversity problems are NP-hard.*

3.1 Utility measurement

In privacy preserving data publishing, it is necessary to balance the conflicting goals, data privacy and utility. There are a number of quality measurements presented in previous studies. Many metrics are utility-based, for example, model accuracy [9, 16] and query quality [15, 38]. They are associated with some specific applications. Three generic metrics have been used in a number of recent works. The discernability metric (DM) was proposed by Bayardo et al. [5] and has been used in [15, 38]. It is defined as follows:

$$\text{DM} = \sum_{\text{QI-group } G} |G|^2$$

where $|G|$ is the size of the QI-group G. The cost of anonymisation is determined by the size of the QI-group. An optimization objective is to minimize weighted discernability cost.

Normalized average QI-group size (CAVG) was proposed by LeFevre et al. [15], and has been used in [38]. It is defined in the following:

$$\text{CAVG} = (\frac{\text{total records}}{\text{total QI-groups}})/(k)$$

The quality of k-anonymisation is measured by the average size of QI-groups produced. An objective is to reduce the normalized average QI-group size.

However, neither the normalized average QI-group size (CAVG) nor the discernability metric (DM) takes the data distribution into account. For this reason we also use the KL-divergence [11], which is described next. In many data mining tasks, we would like to use the published table to estimate the joint distribution of the attributes. Now, given a table T with categorical attributes $A_1, \cdots, A_m$, we can view the data from an m-dimensional distribution F. We can estimate this F with the empirical distribution $\hat{F}$, where $\hat{F}(x_1, \cdots, x_m)$ is the fraction of tuples t in the table such that $t.A_i = x_i$, for $1 \leq i \leq m$. When a generalized version of the table is published, the estimate changes to $\hat{F}^*$ by taking into account the generalizations used to construct the anonymized table T^*(and making the uniformity assumption for all generalized tuples sharing the same attribute values). If the tuple $t = (x_1, \cdots, x_m)$ is generalized to $t^* = (x_1^*, \cdots, x_m^*)$, then $\hat{F}^*(x_1, \cdots, x_m)$ is given by:

$$\hat{F}^*(x_1, \cdots, x_m) = \frac{|\{t^* \in T^*\}|}{|T^*| \times area(t^*)}$$

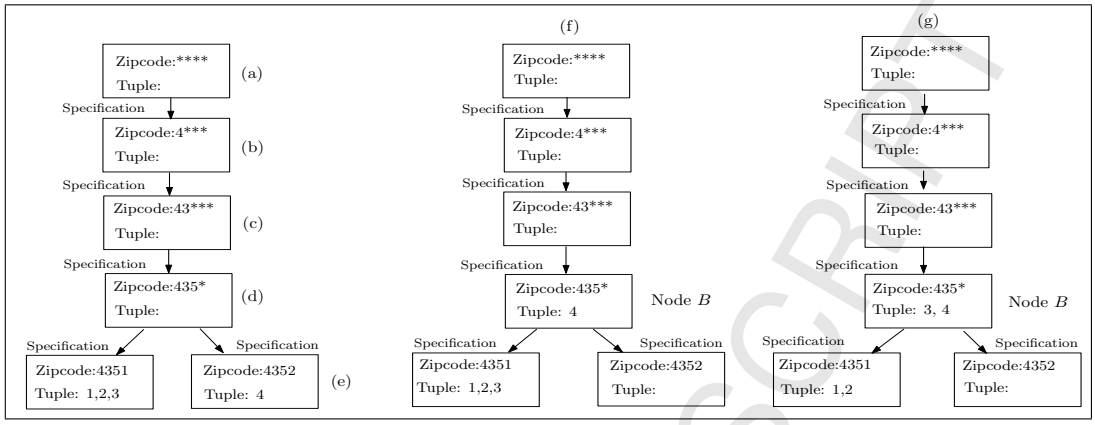


Figure 3: Example illustration of local-recoding algorithm

where, $area(x_1^*, \cdots, x_m^*) = \prod_{i=1}^{m} |\{x_i \in A_i | x_i \text{ is generalized to } x_i^*\}|$.

To quantify the difference between the two distributions $\hat{F}$ and $\hat{F}^*$, we use the Kullback-Leibler divergence (KL-divergence) which is defined as;

$$\text{KL-divergence} = \sum_{x \in A_1 \times \cdots \times A_m} \hat{F}(x) log \frac{\hat{F}(x)}{\hat{F}^*(x)}$$

where $0log0$ is defined to be 0. The KL-divergence is non-negative and is 0 only when the two estimates are identical. In this paper, we use three metrics mentioned above to quantify the information loss of the anonymized data sets.

4 The anonymization algorithms

In this section, we present a top-down approach to tackle the problem of finding an anonymized solution that satisfies distinct (l, α)-diversity, and then we extend it to entropy (l, α)-diversity and recursive (c, l, α)-diversity.

The idea of the algorithm is to first generalize all tuples completely so that, initially, all tuples are generalized into one QI-group. Then, tuples are specialized in iterations. During the specialization, we must maintain distinct (l, α)-diversity. The process continues until we cannot specialize the tuples anymore. For ease of illustration, we present the approach for a quasi-identifier of size 1. The method can be easily extended to handle quasi-identifiers of size greater than 1. The algorithm is shown as in Algorithm 1. Let us illustrate it with the sample data in Table 8(a). Suppose the QI contains Zipcode only. Because there are only

Algorithm 1: The anonymization algorithm ($Distinct(l, \alpha)$)

1. generalize all tuples to their most general form
2. let P be a set containing all these generalized tuples
3. $S \leftarrow \{P\}$; $O \leftarrow \emptyset$.
4. repeat
5. $S' \leftarrow \emptyset$
6. for all $P \in S$ so
7. specialize all tuples in P one level down to form some specialized child nodes.
8. unspecialize the nodes which violate (l, α)-diversity by moving the tuples back to the parent node.
9. if the parent p violates (l, α)-diversity, then
10. unspecialize tuples in the remaining child nodes so that the parent p satisfies (l, α)-diversity
11. for all non-empty branches B of P, do $S' \leftarrow S' \cup \{B\}$
12. $S \leftarrow S'$
13. if P is non-empty then $O \leftarrow O \cup \{P\}$
14. until $S = \emptyset$
15. return O.

No.	Zipcode	Disease
1	4351	HIV
2	4351	Flu
3	4351	HIV
4	4352	Flu

(a)

No.	Zipcode	Disease
1	4351	HIV
2	4351	Flu
3	435*	HIV
4	435*	Flu

(b)

Table 8: (a) Sample data; (b) Generalized table

two distinct sensitive values, we assume that $\alpha = 1$ and $l = 2$. Initially, we generalize all four tuples completely to the most general value Zipcode=**** (Figure 3(a)). Then, we specialize each tuple one level down in the generalization hierarchy. We obtain the branch with Zipcode = 4*** in Figure 3(b). In the next iterations, we obtain the branch with Zipcode = 43** and the branch with Zipcode = 435* in Figure 3(c) and (d), respectively. Next, we further specialize the tuples into the two branches as shown Figure 3(e). Hence the specialization process can be seen as the growth of a tree.

If each leaf node satisfies (l, α)-diversity, the specialization is successful. However, we may encounter some problematic leaf nodes that do not satisfy (l, α)-diversity. Then, all tuples in such leaf nodes will be pushed upwards in the generalization hierarchy. In other words,

those tuples cannot be specialized in this process. They should be kept unspecialized in their parent nodes. For example, in Figure 3(e), the leaf node with Zipcode = 4352 contains only one tuple, which violates (l, α)-diversity when $l = 2$. Thus, we have to move this tuple back to its parent node with Zipcode = 435* (See Figure 3(f)).

After the previous step, we move all tuples in problematic leaf nodes to their parent nodes. However, if the collected tuples in the parent node do not satisfy (l, α)-diversity, we should further move some tuples from other leaf nodes L to the parent node so that the parent node can satisfy (l, α)-diversity while L also maintain the (l, α)-diversity. For instance, in Figure 3(f), the parent node Zipcode = 435* violates (l, α)-diversity when $l = 2$. Thus, we should move one tuples upwards in the node B with Zipcode = 4351 (which satisfies (l, α)-diversity). In this example, we move tuple 3 upwards to the parent node so that both the parent node and the node B satisfy the (l, α)-diversity. Finally, in Figure 3(g), we obtain a data set where Zipcode of tuples 3 and 4 are generalized to 435* and Zipcode of tuples 1 and 2 remains 4351. So the final allocation of tuples in Figure 3(g) is the final distribution of tuples after the specialization. The results can be found in Table 8(b).

The algorithm for generating entropy (l, α)-diverse or recursive (c, l, α)-diverse data set is similar with Algorithm 1. The difference is in the checking criteria of each candidate in the solution space. At the step 8 in Algorithm 1, it tests the (l, α)-diversity property, and in addition to that, we can further test the entropy or recursive conditions to ensure the entropy (l, α)-diversity or recursive (c, l, α)-diversity. We use $Entropy(l, \alpha)$ and $Recursive(l, \alpha)$ to denote the algorithms for entropy (l, α)-diversity and recursive (c, l, α)-diversity.

5 Proof-of-concept experiments

The goals of the experiments are three-fold. First, we study the effect of similarity attacks on the real-life data set by comparing l-diversity with the enhanced (l, α)-diversity models. Second, we evaluate the efficiency of the new proposed models. Third, we investigate the effectiveness of our proposed models in terms of utility preservation.

Experiment setup: We compare four privacy measures, which are l-diversity, distinct (l, α)-diversity, entropy (l, α)-diversity and recursive (c, l, α)-diversity. We compare these privacy measures through evaluations of (1) vulnerability to similarity attacks; (2) efficiency; and (3) data utility. We adopted the publicly available data set, Adult Database, at the UC Irvine Machine Learning Repository[3], which has become the benchmark of this field [14, 23, 9]. We used a configuration similar to [14, 23] by eliminating the records with unknown values. The resulting data set contains 45,222 tuples. Seven of the attributes were chosen as the quasi-

[3]available at `www.ics.uci.edu/-mlearn/MLRepository.html`

Attribute	Type	Distinct values	Height
Age	Numeric	74	5
Workclass	Categorical	8	3
Education	Categorical	16	4
Country	Categorical	41	3
Marital Status	Categorical	7	3
Race	Categorical	5	3
Gender	Categorical	2	2
Health Condition	Sensitive	8	–

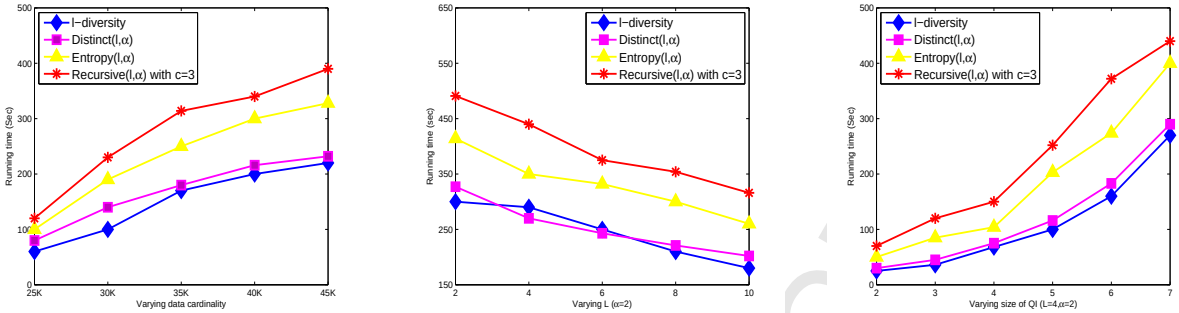
Table 9: Description of Quasi-identifier

identifier. We add a column with sensitive values called "Health Condition" consisting of {HIV, Cancer, Phthisis, Hepatitis, Obesity, Asthma, Flu, Indigestion} to the Adult data and randomly assign one sensitive value to each record. Table 9 provides a brief description of the data including the attributes we used, the type of each attribute data, the number of distinct values for each attribute, and the height of the generalization hierarchy for each attribute. We divide the 8 values of the Health Condition attribute into four pre-defined equal-size categories, based on the confidentiality of the values (See Table 4). In this paper, the weight of each category is evaluated by using the sensitivity weight function with $\beta = 2$ defined in Section 3. We implement $Distinc(l, \alpha)$, $Entropy(l, \alpha)$ and $Recursive(l, \alpha)$ algorithms for the distinct (l, α)-diversity, entropy (l, α)-diversity and recursive (c, l, α)-diversity.

Similarity attack: We use the first 7 attributes in Table 9 as the quasi-identifier and treat Health Condition as the sensitive attribute, and divide the eight values of the Health Condition into four categories groups shown in Table 4. Any QI-group that has all values falling in one category is viewed as vulnerable to the similarity attacks. We use the modified Incognito [14] to generate 4-diverse table. In the anonymized table, a total of 1570 tuples can be inferred about their sensitive value categories. The results show that similarity attacks present serious privacy risks to l-diverse tables on real data. We also generate the anonymized table satisfying the distinct (4,2)-diversity, and entropy (4,2)-diversity, and both tables do not contain tuples that are vulnerable to similarity attacks. This shows that both the distinct (l, α) diversity and entropy (l, α)-diversity provide better privacy protection against similarity attacks.

Efficiency: In this set of experiments, we compare the running time of the algorithms for finding l-diverse, distinct (l, α)-diverse ($Distinct(l, \alpha)$), Entropy (l, α)-diverse ($Entropy(l, \alpha)$) and recursive (l, α)-diverse ($Recursive(l, \alpha)$) data sets. The results are shown in Figure 4.

Data used for Figure 4(a) is generated by re-sampling the Adult data sets while varying the cardinality of data from 25K to 45K. We evaluate the running time for all privacy measures



(a) (b) (c)

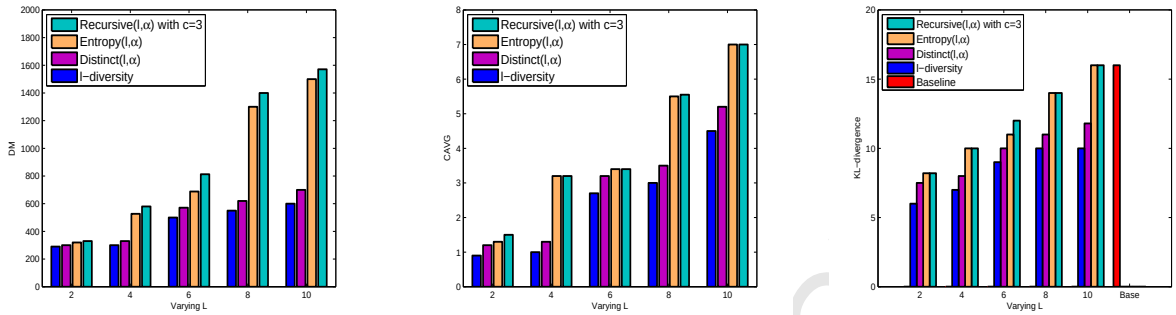
Figure 4: Performance comparisons of four privacy measures by: (a) varying data cardinality; (b) varying l; (c) varying the size of QI.

with default setting $l = 4, \alpha = 2$. From Figure 4(a) we can see, the execution time for all the anonymization algorithms is ascending with the increasing data percentage. This is because as the percentage of data increases, the computation cost increases too. The result is expected since the overhead is increased with more dimensions.

Next, we evaluate how the parameter l affects the cost of computing. Data set used for this set of experiments is the whole Adult data and we evaluate by varying l. Setting $\alpha = 2$, Figure 4(b) displays the results of execution time by varying l from 2 to 10. The cost drops as l grows, because the larger the l is, the more chance each QI-group has more distinct sensitive values, which makes it easier to meet the privacy requirements, therefore allowing our algorithms to terminate earlier.

Finally, we evaluate the effect of the size of QI attributes on the computation overhead. We vary the size of the set of quasi-identifier attributes from 2 to 7. A QI attribute set of size j consists of the first $j - 1$ attributes listed in Table 9 and Health Condition as the sensitive attribute. We measured the time taken to return all 4-diverse, distinct (4,2)-diverse, entropy (4,2)-diverse and recursive (3,4,2)-diverse data sets. As we can see from Figure 4(c), the running time for finding the (4,2)-diversity is always less than the entropy (4,2)-diversity and recursive (3,4,2)-diversity. This is because higher privacy requirements need more computation cost. Also, finding the 4-diverse data set has the similar computation cost with searching for the (4,2)-diverse data set for the adult database, which makes the enhanced (l, α)-diversity models practical.

From these evaluations, the running times for enhanced (l, α)-diversity models are fast enough for them to be used in practice, and more important is the enhanced (l, α)-diversity models can effectively prevent from similarity attack.



(a) (b) (c)

Figure 5: Data utility comparison by varying l vs. : (a) discernability metric (DM); (b) normalized average QI-group size (CAVG); (c) vs. KL-divergence

Data utility: Having verifying the efficiency of our technique, we proceed to test its effectiveness. The utility is measured by three metrics introduced in Section 3.1. Figure 5(a) displays the comparison of different methods when varying l on Adult database with discernability metric (DM). We can see that for the variable l, the produced anonymizaed tables satisfying (l, α)-diversity preserves the similar data utility compared with l-diversity model, and it maintains better data utility than the other two enhanced privacy principles. This is because the requirements of the other two principles are more strict, which may require more generalization operations, and make the discernability metric (DM) higher than the distinct (l, α)-diversity and l-diversity. Figure 5(b) reports the results with regard to the normalized average QI-group size (CAVG). We can see that for a smaller l, the quality of the data is better preserved for three variants of (l, α)-diversity models, but for a larger value of l, the entropy (l, α)-diversity and recursive (l, α)-diversity models produce the anonymized data sets with less utility than the distinct (l, α)-diversity principle.

In Figure 5(c), we compare l-diverse, distinct (l, α)-diverse, entropy (l, α)-diverse and recursive (c, l, α)-diverse tables using the KL-divergence utility metric. We wish to publish a table from which the joint distribution $Q \times S$ can be estimated, where $S =$ Health Condition and Q is the first seven attributes in Table 9. The baseline in Figure 5(c) corresponds to the KL-divergence for the table where all the attributes in Q were completely suppressed (thus the resulting table had only one attribute-the sensitive attribute). This table represents the least useful anonymized table that can be published. The rest of the bars correspond to the KL-divergence to the best distinct (l, α)-diversity, entropy (l, α)-diversity and recursive $(3, l, \alpha)$-diverse tables, respectively for $l = 2, 4, 6, 8, 10$, $\alpha = 2$. In the experiments run on the full Adults data set, we see that the KL-divergence to the best (l, α)-diverse table (entropy

or recursive) is very close to the KL-divergence to the best distinct (l, α)-diverse table when $l = 2, 4, 6$. As expected, for larger values of l, the utility of (l, α)-diverse tables is lower, and the best tables for the entropy and recursive variants of the (l, α)-diversity models often have similar utility. Hence, for $l = 8, 10$ the best tables were very close to the baseline. For $l = 6$, the recursive definition performs better than the entropy definition since recursive $(3, l, \alpha)$-diversity allows for more skew in the sensitive attribute.

6 Related Work

The problem of information disclosure has been studied extensively in the framework of statistical databases. A number of information disclosure limitation techniques have been designed for data publishing, including Sampling, Cell Suppression, Rounding, and Data Swapping and Perturbation. These techniques, however, insert noise to the data. Samarati and Sweeney [26, 31, 30] introduced the k-anonymity model. Since then, there has been a large amount of research work on this topic.

The first category of work aims at devising privacy requirements. The k-anonymity model [26, 31, 30] assumes that the adversary has access to some publicly available databases (e.g., a vote registration list) and the adversary knows who is and who is not in the table. A few subsequent works [23, 34, 36, 18, 19, 21] recognize that the adversary also has knowledge of the distribution of the sensitive attribute in each group.

Privacy-preserving data publishing has been extensively studied in several other aspects. First, background knowledge presents additional challenges in defining privacy requirements. Several recent studies [7, 22, 20, 19] have aimed at modeling and integrating background knowledge in data anonymization. Second, several works [6, 39, 28] considered continual data publishing, i.e., re-publication of the data after it has been updated. A m-invariance is one of the representative models [39]. The basic idea is to keep unchanged the set of sensitive attribute values in the group that a tuple belongs to even though the tuple may be put into different groups in different versions of the microdata. Our proposed enhanced (l, α)-diversity models can also be extended for dynamic microdata. Nergiz et al. [17] proposed σ-presence to prevent membership disclosure, which is different from identity/attribute disclosure. Wong et al. [37] showed that knowledge of the anonymization algorithm for data publishing can leak extra sensitive information. Recently, Koudas et al. [25] designed anonymization schemes that disguise the distribution of sensitive attributes of microdata, which allows accurately answer aggregate queries. The designed schemes also support flexible, user-defined tradeoff between privacy and data utility.

We want to emphasize that l-diversity is still a useful measure for data publishing. l-diversity and our enhanced measures make different assumptions about the adversary. l-

diversity assumes an adversity who has knowledge of the form "Someone does (not) have some kind of disease", while our enhanced measures further consider an adversary who knows the distributional information of the sensitive attributes. Our goal is to propose an alternative technique for data publishing that remedies the limitations of l-diversity in some applications.

Most anonymization solutions adopt generalization [1, 5, 9, 14, 15, 16, 26, 31] and bucketization [22, 41]. In this paper, we use the Incognito algorithm [14] to implement l-diversity [23]. There are several anonymization techniques, clustering [3], marginal's releasing [12, 29], data perturbation [40] and micro-aggregation [8]. On the theoretical side, optimal k-anonymity has been proved to be NP-hard for $k \geq 3$ in [24, 2, 27], and approximation algorithms for finding the anonymization that suppresses the fewest cells have been proposed in [24, 2]. The anonymization method we used in this paper is generalization.

7 Conclusion and future work

l-diversity is a novel property that, when satisfied by microdata, can help increase the privacy of the respondents whose data is being used. However, as shown in the paper, to some extent this property is not enough for protecting sensitive attributes. In this paper, we proposed a family of enhanced (l, α)-diversity models against sensitive attribute disclosures. We theoretically analyzed the hardness of this series of problems, and developed efficient algorithms to deal with them. Our extensive experiments show that our proposed methods are effective and practical in real-world applications.

This work also initiates several directions for future investigation. For example, in this article, we focused on the case where there is a single sensitive attribute; extending our technique to multiple sensitive attributes is an interesting topic. Another direction concerns the comparison with t-closeness [18], and some experimental evaluations can be done to compare the two privacy principles. Finally, it would be useful to study how the anonymized data sets can be utilized for discovering complex patterns, perhaps through minimization of specialized metrics for quantifying information loss.

Acknowledgement

We would like to thank for reviewers' useful comments on improving the quality of the paper. This research is supported by Australian Research Council (ARC) grant DP0774450 and DP0663414.

References

[1] C. Aggarwal, On k-Anonymity and the Curse of Dimensionality, Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp. 901-909, 2005

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas and A. Zhu. Anonymizing tables. *In Proc. of the 10th International Conference on Database Theory (ICDT'05)*, pp. 246-258, Edinburgh, Scotland.

[3] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, Achieving Anonymity via Clustering, Proc. of the ACM Symp. on Principles of Database Systems (PODS), pp. 153-162, 2006.

[4] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *J. Am. Stat. Assoc.*, pages 10-28, 1986

[5] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymity. *In Proceedings of the 21st International Conference on Data Engineering (ICDE)*, 2005.

[6] J. W. Byun, Y. Sohn, E. Bertino, and N. Li, Secure Anonymization for Incremental Datasets, Secure Data Management (SDM), pp. 4863, 2006

[7] B. C. Chen, K. LeFevre, and R. Ramakrishnan, Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge, Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp. 770-781, 2007.

[8] J. Domingo-Ferrer and V. Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery. v11. 195-212.

[9] B. Fung, K. Wang, P. Yu. Top-down specialization for information and privacy preservation. *In Proc. of the 21st International Conference on Data Engineering (ICDE'05)*, Tokyo, Japan.

[10] M. R. Garey, D. S. Johnson. Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco. Freeman, 1979.

[11] S. Kullback and R. Leibler. On information and sufficiency. Annals of Mathematical Statistics, 22(1):79-86, 1951.

[12] D. Kifer and J. Gehrke, Injecting Utility into Anonymized Datasets, Proc. of the ACM SIGMOD Intl Conf. on Management of Data (SIGMOD), pp. 217-228, 2006

[13] D. Lambert. Measure of disclosure risk and harm. *Journal of Official Statistics*, vol 9, 1993, pp. 313-331.

[14] K. LeFevre, D. DeWitt and R. Ramakrishnan. Incognito: Efficient Full-Domain k-Anonymity. *In ACM SIGMOD International Conference on Management of Data*, June 2005.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In ICDE'06: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), page 25, Washington, DC, USA, 2006. IEEE Computer Society.

[16] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware Anonymisation. In KDD'06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 277-286, Philadelphia, PA, USA, 2006. ACM Press.

[17] M. E. Nergiz, M. Atzori, C. Clifton, Hiding the Presence of Individuals from Shared Databases, Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 665-676, 2007.

[18] N. Li, T. Li, S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *ICDE 2007*: 106-115

[19] T. Li, N. Li, and J. Zhang. Modeling and Integrating Background Knowledge in Data Anonymization. ICDE, 2009.

[20] T. Li and N. Li, Injector: Mining Background Knowledge for Data Anonymization. In Proc. Int'l Conf. Data Engineering (ICDE), 2008.

[21] T. Li, N. Li. On the Tradeoff Between Privacy and Utility in Data Publishing. To appear in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2009.

[22] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, Worst-Case Background Knowledge for Privacy-Preserving Data Publishing, Proc. Intl Conf. Data Engineering (ICDE), pp. 126-135, 2007

[23] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In ICDE, 2006.

[24] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. *In Proc. of the 23rd ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*, pp. 223-228, Paris, France, 2004.

[25] N. Koudas, D. Srivastava, T. Yu and Q. Zhang. Distribution-based Microdata Anonymization. to appear in in the 35th International Conference on Very Large Data Bases (VLDB), 2009.

[26] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010-1027. 2001

[27] X. Sun, H. Wang and J. Li. On the Complexity of Restricted k-anonymity Problem. APWeb 2008: 287-296

[28] X. Sun, H. Wang and J. Li. L-Diversity Based Dynamic Update for Large Time-Evolving Microdata. Australasian Conference on Artificial Intelligence 2008: 461-469

[29] X. Sun, H. Wang and J. Li. Injecting purpose and trust into data anonymization. to appear in CIKM 2009.

[30] L. Sweeney.: Achieving k-anonymity Privacy Protection Using Generalization and Suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*, 10(5) pp. 571-588, 2002.

[31] L. Sweeney. k-anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty Fuzziness Knowledge-based Systems*, 10(5), pp 557-570, 2002

[32] T. M. Traian and V. Bindu, Privacy Protection: p-Sensitive k-Anonymity Property *International Workshop of Privacy Data Management (PDM2006), In Conjunction with 22th International Conference of Data Engineering (ICDE)*, Atlanta, 2006.

[33] T. M. Traian, A. Campan and P. Meyer. Generating Microdata with P-sensitive k-anonymity Property. *SDM 2007*: 124-141

[34] K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In ICDM05, 2005

[35] W. E. Winkler. Advanced Methods for Record Linkage, *Proceedings of the Section on Survey Research Methods, American Statistical Society*, 467-472

[36] R. Wong, J. Li, A. Fu, K. Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. *KDD 2006*: 754-759.

[37] R. Wong, A. Fu, K. Wang, and J. Pei, Minimality Attack in Privacy Preserving Data Publishing, Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp. 543-554, 2007.

[38] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. C. Fu. Utility-based Anonymisation using local recoding. In KDD'06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 785-790, Philadelphia, PA, USA, 2006. ACM Press.

[39] X. Xiao and Y. Tao, m-invariance: Towards Privacy Preserving Republication of Dynamic Datasets, Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 689-700, 2007.

[40] X. Xiao and Y. Tao, Personalized Privacy Preservation, Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 229-240, 2006

[41] X. Xiao and Y. Tao, Anatomy: simple and effective privacy preservation, Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp. 139-150, 2006.